Numériser le patrimoine I: standards et bonnes pratiques

Du texte à l'objet: Décrire et échanger les données

Simon Gabay

Genève

Remarques introductives

In principio erat verbum

- Importance de la linguistique computationnelle dans les humanités numériques (compter les mots)
- La TEI fondée par l'Association for Computers and the Humanities, l'Association for Computational Linguistics et l'Association for Literary and Linguistic Computing
- D'où l'importance de l'étude des textes avant les autres choses (objets, musique, films…)

Description et échange des données

- Importance des institutions patrimoniales (musées, bibliothèques, archives)
- Les systèmes d'échange synthétisent les données essentielles à la description
- XML est le moyen privilégié de l'échange de données – et il est lisible par l'être humain
- Autant de raison de s'attarder sur ces formats, entre autres pour des raisons pédagogiques
- Il existe évidemment des systèmes bien plus complexes…

De nouveau la TEI

`<MsDesc>`

`<MsDesc>` permet de décrire le manuscrit

- `<msIdentifier>` pour la cote
- `<author>` pour l'auteur
- `<docDate>` pour la date
- `<support>` pour la description du matériaux (parchemin, vélin...)
- `<extent>` pour le format (taille, longueur...)
- `<condition>` pour son état de conservation
- La description peut être extrêmement complexe (mains, enluminures, sceaux, filigranes)
- Description de manuscrit: _Antiphonarium Lausannense. De Sanctis, pars hiemalis. Officium B.M.V. Commune Sanctorum: sur www.e-codices.ch

`<MsDesc>` +

- "Détournement" (ou plus précisément "changement de sémantisme") de
- Bibliographie matérielle, pour les catalogues de livres (anciens)
- Pour décrire le support des inscriptions épigraphiques
- Description d'épigraphie (cf. ISic0298)

Dublin core

- *Dublin Core Metadata Initiative* (DCMI)
- 1995, Dublin (Ohio, pas Irlande)
- Décrire des documents de manière simple et standardisée
  En deux parties
- *Dublin Core element set*: quinze propriétés
- *Dublin Core metadata terms*: d'autres propriétés supplémentaires
  Dublin Core element set en deux types
- Éléments de métadonnées Dublin Core
- Autres

Dublin Core element set: Métadonnées

Nom	Description
Title	Nom donné à la ressource
Creator	Nom de la personne responsable de la création de la ressource
Subject	Thème du contenu
Description	Présentation du contenu
Date	Date de création
Language	Langue du contenu intellectuel
Relation	Référence à une ressource apparentée
Coverage	Couverture spatio-temporelle
Rights	Informations sur les droits associés

DCMI element set

Nom	Description
Publisher	Organisme de diffusion
Contributor	Personne responsable de contributions au contenu
Type	Nature ou genre
Format	Manifestation physique ou numérique
Identifier	Référence univoque dans un contexte donné (URI, ISBN)
Source	Référence dont la ressource décrite est dérivée (URI)
...	...

Metadata Terms (extension de l'*element set*)

- dateCopyrighted
- rightsHolder
- created
- issued
- provenance
- isPartOf
- isVersionOf
- hasVersion
- tableOfContents

Entre vocabulaire et langage

- Dublin core est un vocabulaire du web sémantique
- Il utilisé pour exprimer les données dans un modèle RDF (*Ressource description framework*)
- Il peut être exprimé avec une syntaxe XML (`.xml`)
- Il peut être exprimé avec une syntaxe Turtle (`.ttl`)
- Il peut être exprimé avec une syntaxe N-Triples (`.nt`)

Plus loin que DC

- *MAchine-Readable Cataloging* (MARC)
- *Metadata Object Description Schema* (MODS, entre DC et MARC)
- *Metadata Encoding and Transmission Standard* (METS)

Echanger les données

- Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
  -> Exemple d'e-codices
- SRU=Search/Retrieve via URL
  -> Exemple de swissbib

LIDO

Lightweight Information Describing Objects

- C’est un format d’échange de données
- Il permet de décrire les objets et les ressources numériques (images, textes, sons, vidéos)
- 14 groupes d’informations, dont 3 sont obligatoires
- 5 types de groupes d’information

LIDO 1: classification

1. **Object/Work type** (classification)
2. Classification (style, forme, âge…)

LIDO 2: événements

3. Event set (création, exposition. . . On y reviendra)

LIDO 3: relations

4. Subject set (objet, bâtiments, personnes dans l’œuvre)
5. Related Works

LIDO 4: identification

6. **Title/Name**
7. Inscriptions (transcription et ou description)
8. Repository/location (institution et numéro d’inventaire)
9. State/Edition
10. Object description
11. Measurements

LIDO 5: Administration

12. Rights
13. **Record**
14. Ressources

Autres formats pour les musées

- museumdat (www.museumdat.org)
- SPECTRUM XML (http://www.collectionstrust.org.uk/spectrum)
- CIDOC-CRM (http://www.cidoc-crm.org/)

LIDO et CIDOC-CRM

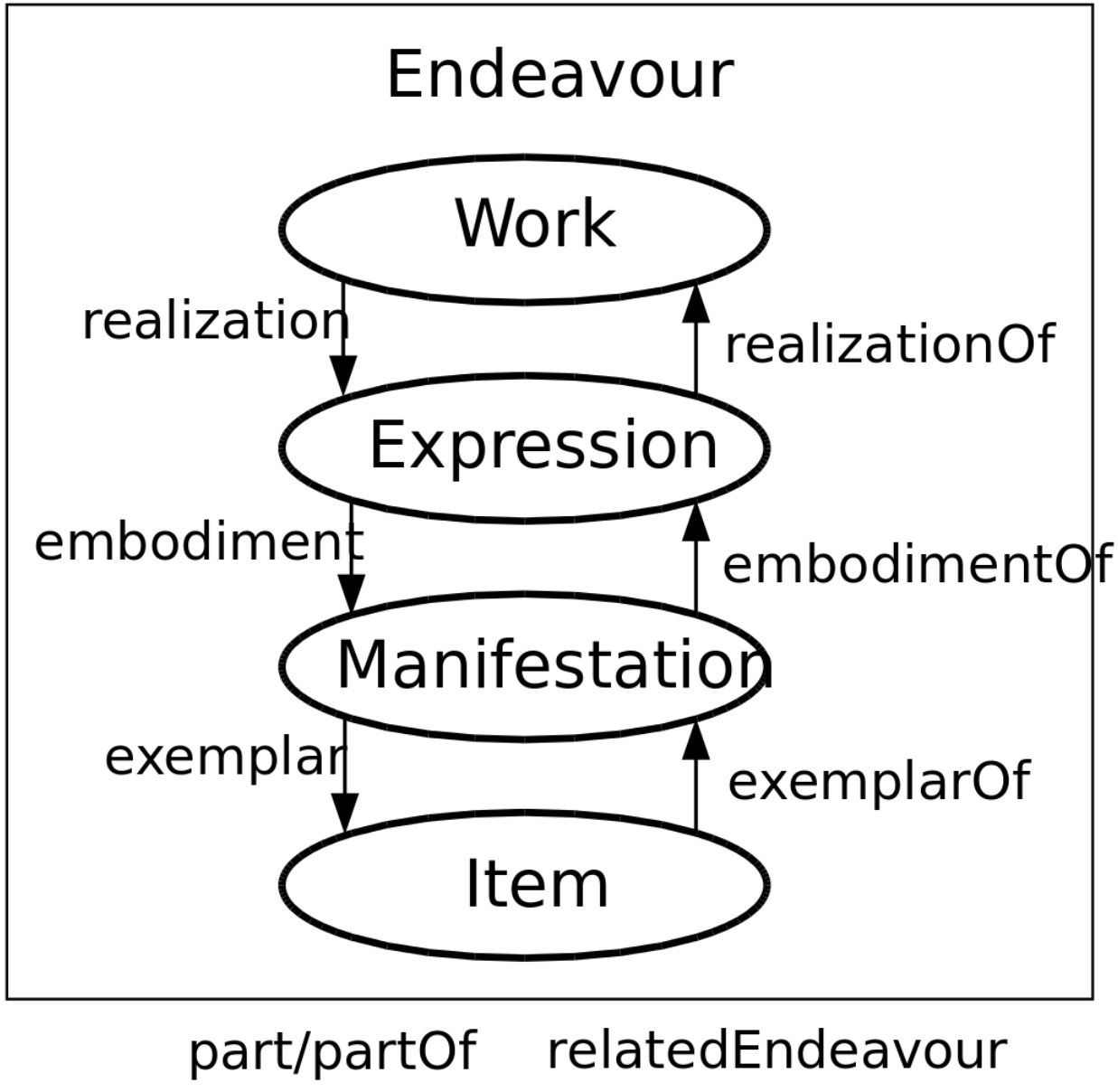
Lien avec CIDOC-CRM (*Conceptual reference model*)

- LIDO est un format de description (comme DC)
- CIDOC CRM est un modèle conceptuel (comme FRBR)
- Ce modèle de données est "orienté événement"

Ce modèle "orienté événement" décrit les événements de la vie d'un objet pour décrire ce dernier *via*:

- un/des agent(s)
- une date ou un intervalle dans le temps
- un lieu
- un type d'événement

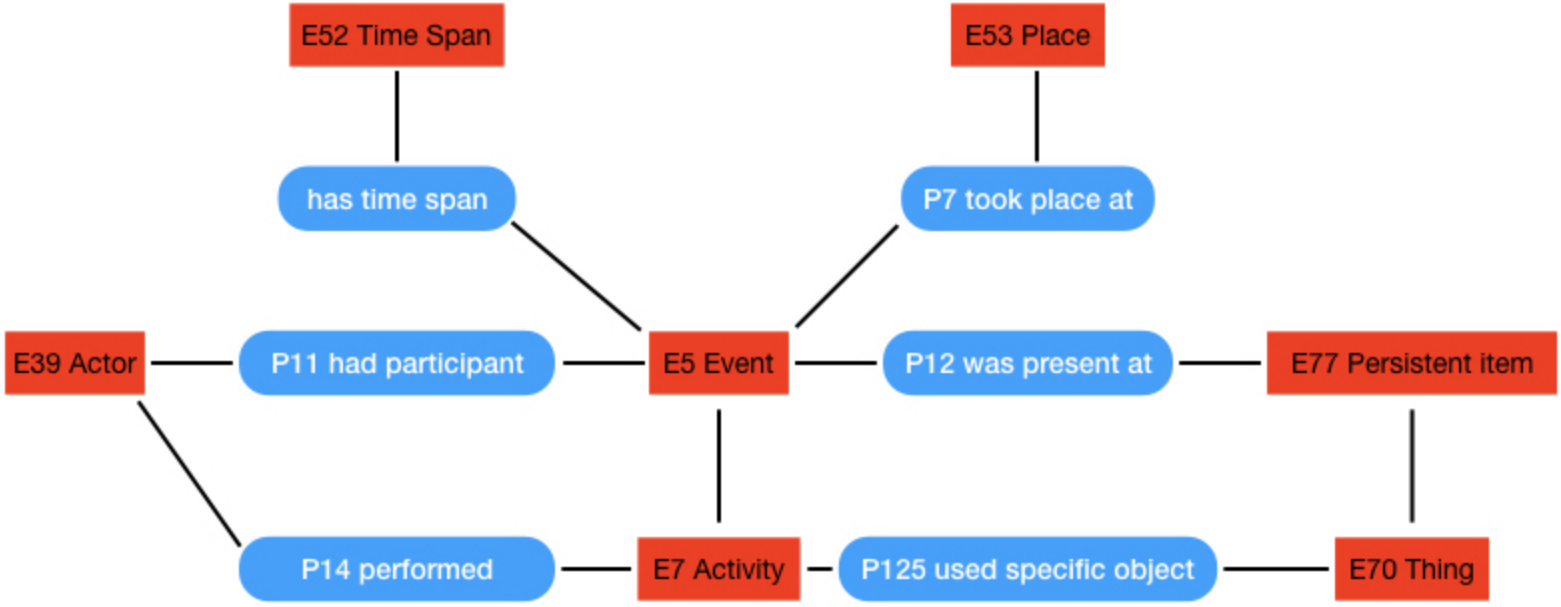
FRBR


Endeavour
Work
realization
realizationOf
Expression
embodiment
embodimentOf
Manifestation
exemplar
exemplarOf
Item
part/partOf
relatedEndeavour

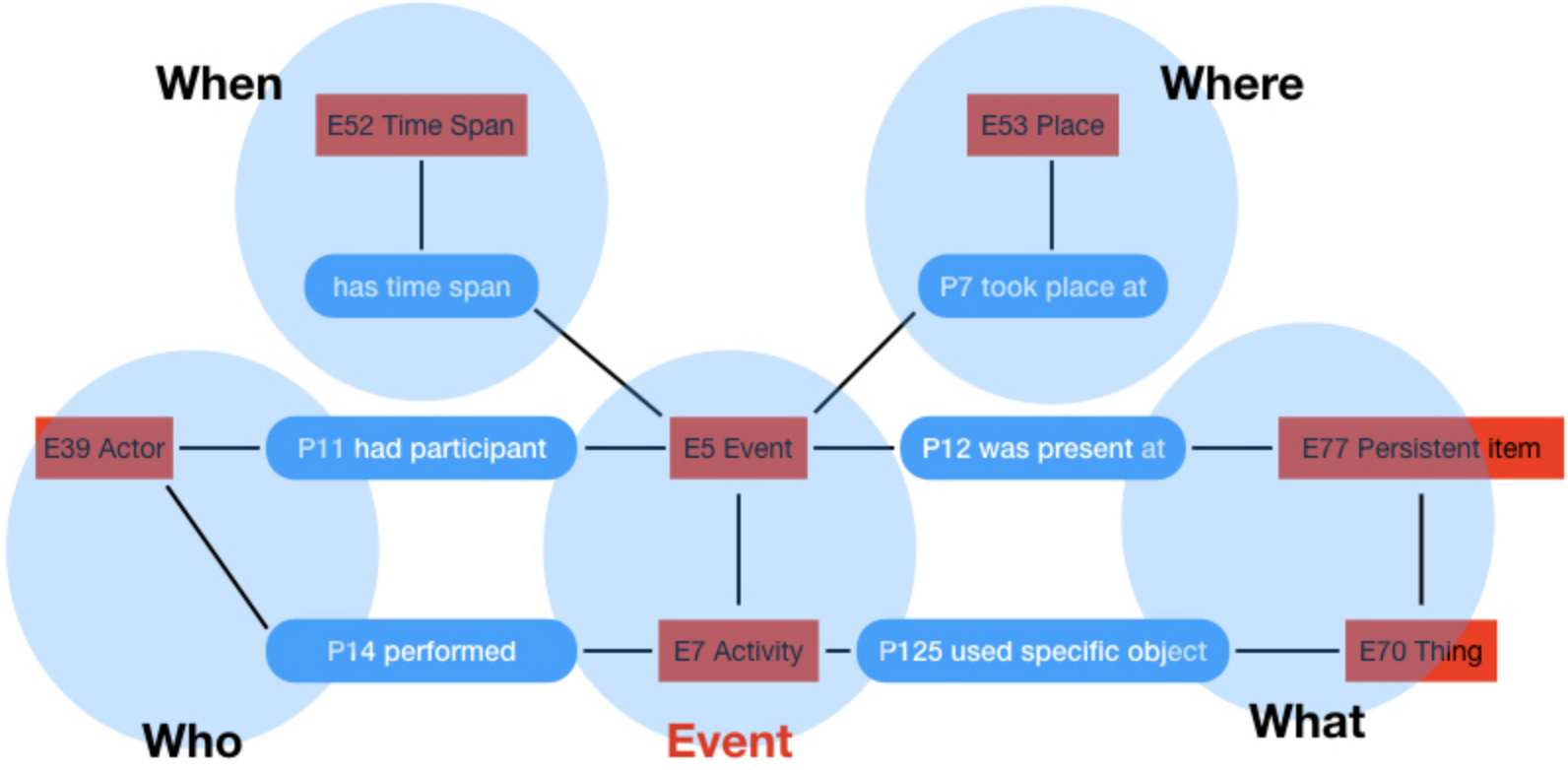
Types d’événements

- Creation
- Modification
- Part addition
- Part removal
- Excavation
- Acquisition
- Finding
- Exhibition
- Move
- Restoration
- Loss
- Destruction

CIDOC-CRM simplifié



Les quatre W (*who*, *what*, *when*, *where*)



Exemple: Le *Monument à Balzac* de Rodin